Supplementary figures

Figure legends

Figure S1:

Comparison of the annotations inferred on trees built with Probcons+BMGE+RAxML (left panel, same as article Fig. 2), Probcons+BMGE+Phylobayes (middle panel) or MAFFT-EINSI+Guidance+RAxML (right panel). The annotations obtained on the Probcons+BMGE+PhyML (UL3 model) maximum-likelihood trees are also reported (left panel, “UL3” column). The order of LECA clades is the same in all panels. LECA clade identifiers and definitions are reported on the right of each line. White lines in Phylobayes and MAFFT-Guidance panels correspond to LECA clades for which the support for the monophyly of eukaryotes was under 50% using those alternative approaches, or (for Phylobayes only) to LECA clades whose MCMC chains were stopped before convergence was achieved (i.e. maximum bipartition discrepancy <0.3). The correlation between all three approaches is excellent, except the topologies sampled with the Bayesian method are generally distributed more narrowly around the ML topology, as expected [1]. For each cluster, the species-presence profile is summarized on the left. Each column corresponds to a high-order prokaryotic group. A small colored circle indicates that a few species of the group are present ; a large colored circle indicates that more than half the species of the group are present (the thresholds given in article table 1 are used) ; a black ring indicates that a clade comprising at least half the species of the group exists in the ML tree of the cluster.

1. Douady CJ, Delsuc F, Boucher Y, Doolittle WF, Douzery EJP: **Comparison of Bayesian and Maximum Likelihood Bootstrap Measures of Phylogenetic Reliability**. *Mol Biol Evol* 2003, **20**:248–254.

Figure S2:

Phylogenies of the LECA gene HBG298928_1 ("Alpha/beta hydrolase fold protein") using 183 (left) or 882 (right) prokaryotic genomes. This figure illustrates one of the subtle issues caused by sampling. At first this gene was labeled "actinobacteria-related" on the basis of the 183-prokaryotic-genomes dataset, as all its homologs were actinobacterial. It then appeared using the 882-prokaryotic-genomes dataset that it was also present at low frequencies in at least three orders of gammaproteobacteria. In addition, depending on where the larger tree was rooted, the sister group of Eukarya was Actinobacteria, Gammaproteobacteria, or both. Thus an important part of the molecular history of this family of homologs was missed, what rendered the initial annotation questionable.

Figure S3:

Phylogeny of the LECA gene HBG487932_1 ("Cytochrome b"). This tree illustrates the limits of considering only the support of the node at the base of the stem of Eukaryotes. For this LECA clade, the NBS support was 7%, even though the tree is rather well resolved. In contrast, the SGS score was 68%.

Figure S4:

LECA clades displayed according to the first and second levels of the KEGG ORTHOLOGY ontology. Some LECA clades may appear in several categories.

Figure S5:

Comparison of the annotations of LECA clade positions based on the “reference” configurations (article Fig. 1), a “relaxed” criterion or a “naive” one. (A) Schematic diagrams for the criteria. The relaxed criterion was similar to the reference one except eukaryotes are allowed to branch as a sister group of the putatively related prokaryotic sequences (when the reference one required that they branched among them). With the naive criterion, a relationship was inferred whenever eukaryotes have a taxonomically homogeneous sister clade, even when this clade was made of only one sequence (e.g. main-text Fig. 1C). (B) Annotations obtained using the naive (left), relaxed (middle) or reference (right) criteria. The figure is to be read like main-text Fig. 2. The LECA clades appear in the same order in the three panels. The sorting is based on the left panel (note that the rows of the right panel are the same as those of article Fig. 2, but resorted).

The greatest difference was between the naive and relaxed criteria, that is when taxonomic representativeness criteria were introduced. Most often, the relaxed and reference criteria differed only quantitatively. The reference criterion was adopted because it was immune to rooting issues.

Interestingly, the ability of the method to detect putatively alphaproteobacterial genes varied little with the criterion used. Indeed, the number of candidate alphaproteobacteria-related LECA clades using the naive and reference criteria were respectively 46 ($\>$50\% of alphaproteobacteria-related bootstrap replicates) and 41 ($\>$5\% of alphaproteobacteria-related replicates representing more than 80\% of all replicates annotated to a specific taxonomic group). The reason for this result is that configurations were much more specific, so that even LECA clades for which the alphaproteobacteria-related configuration appeared in a few replicates could be considered candidates.

Figure S1

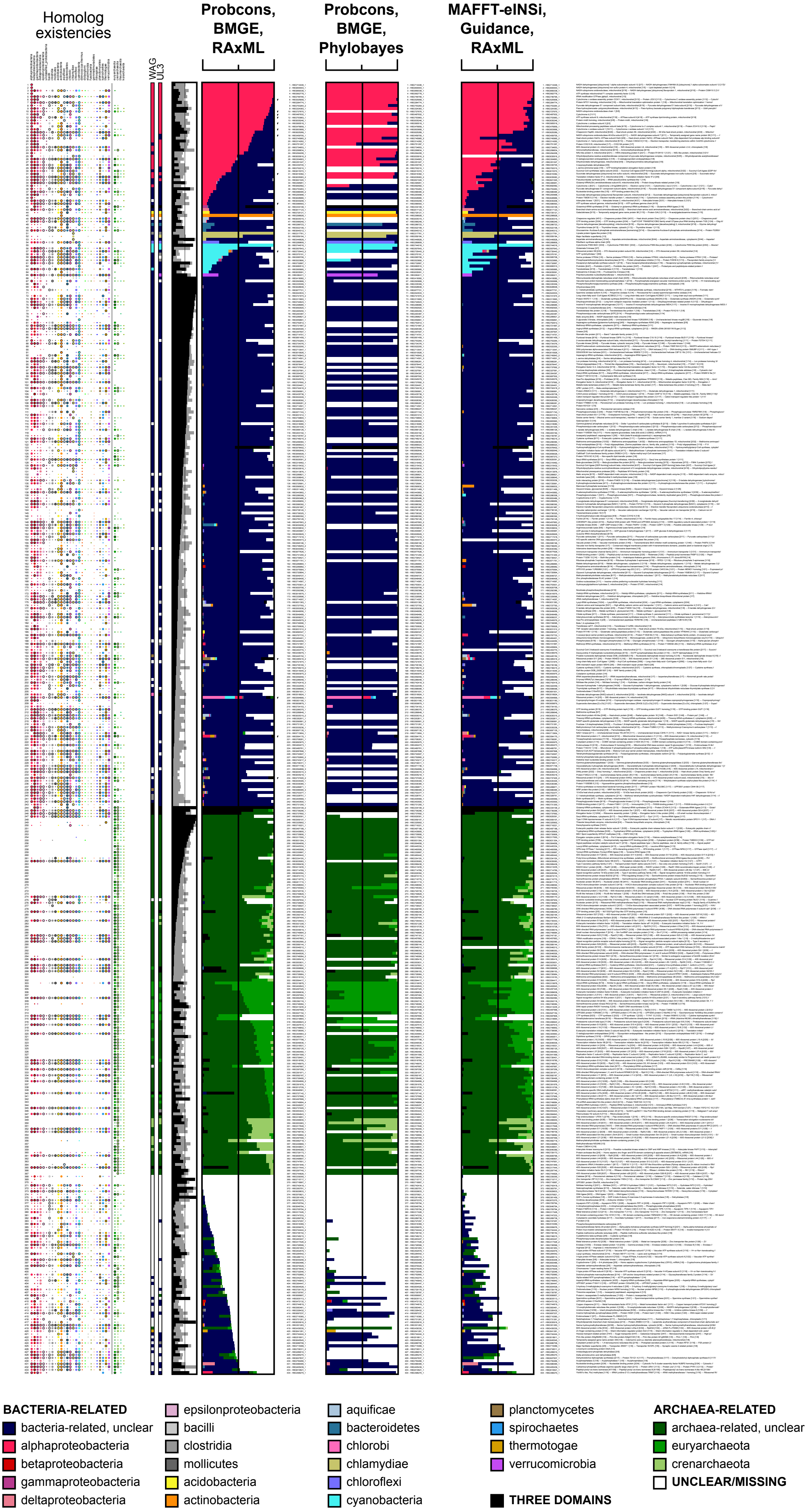
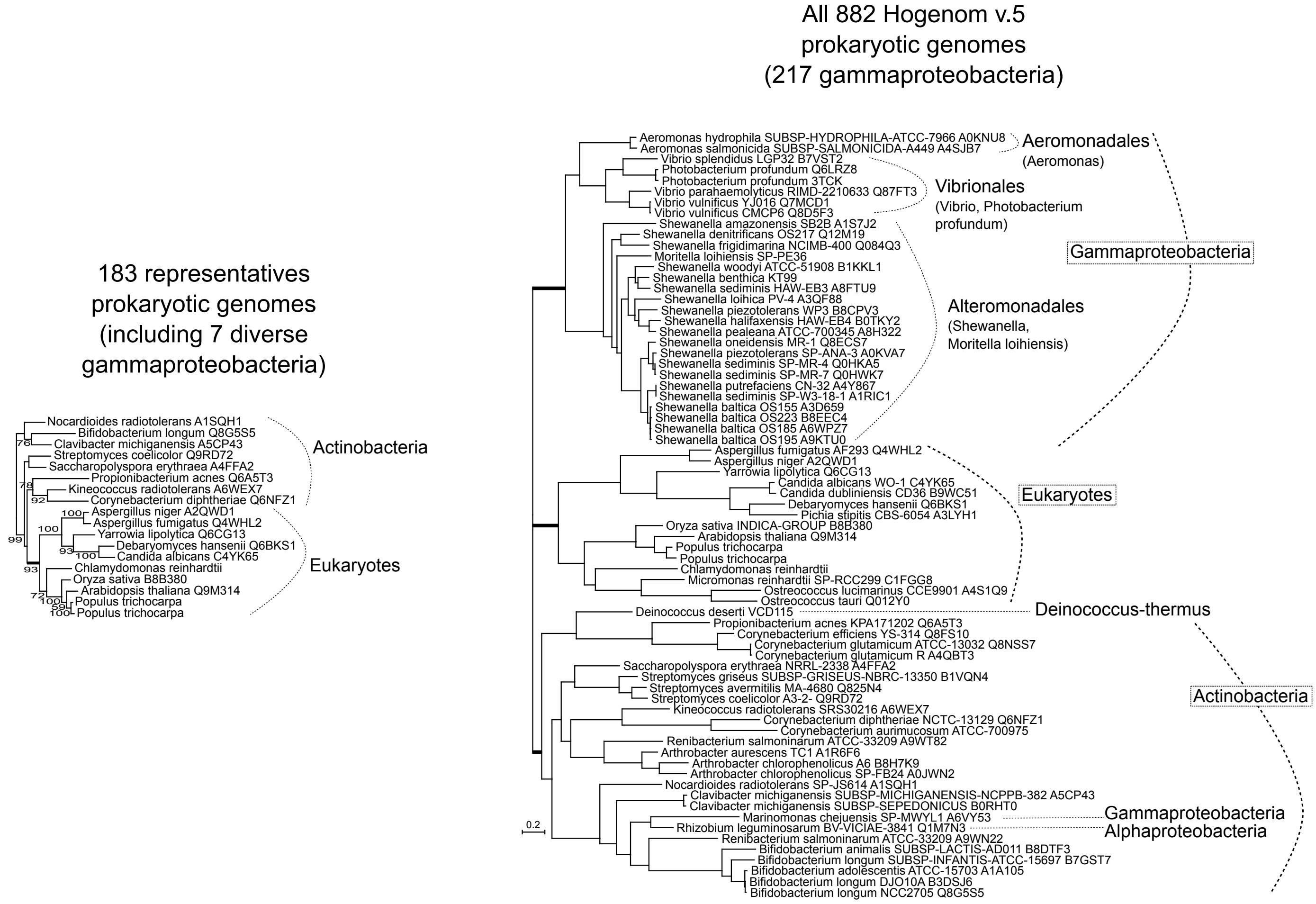

Homolog existencies
WAG
UL3
Probcons, BMGE, RAxML
Probcons, BMGE, Phylobayes
MAFFT-eINSi, Guidance, RAxML
BACTERIA-RELATED
bacteria-related, unclear
alphaproteobacteria
betaproteobacteria
gammaproteobacteria
deltaproteobacteria
epsilonproteobacteria
bacilli
clostridia
mollicutes
acidobacteria
actinobacteria
aquificae
bacteroidetes
chlorobi
chlamydiae
chloroflexi
cyanobacteria
planctomycetes
spirochaetes
thermotogae
verrucomicrobia
THREE DOMAINS
ARCHAEA-RELATED
archaea-related, unclear
euryarchaeota
crenarchaeota
UNCLEAR/MISSING

Figure S2

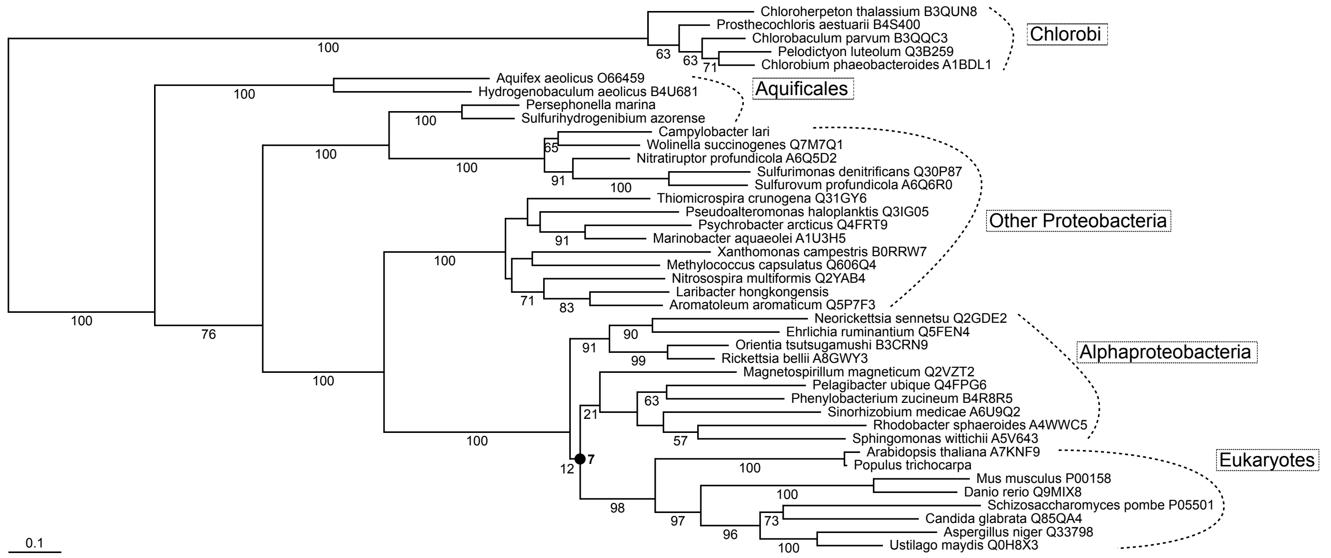


All 882 Hogenom v.5 prokaryotic genomes (217 gammaproteobacteria)

183 representatives prokaryotic genomes (including 7 diverse gammaproteobacteria)

Nocardioides radiotolerans A1SQH1
Bifidobacterium longum Q8G5S5
76
Clavibacter michiganensis A5CP43
Streptomyces coelicolor Q9RD72
Saccharopolyspora erythraea A4FFA2
Propionibacterium acnes Q6A5T3
78
Kineococcus radiotolerans A6WEX7
92
Corynebacterium diphtheriae Q6NFZ1
Actinobacteria
100
Aspergillus niger A2QWD1
Aspergillus fumigatus Q4WHL2
100
Yarrowia lipolytica Q6CG13
99
Debaryomyces hansenii Q6BKS1
93
100
Candida albicans C4YK65
93
Chlamydomonas reinhardtii
Oryza sativa B8B380
72
Arabidopsis thaliana Q9M314
100
59
Populus trichocarpa
100
Populus trichocarpa
Eukaryotes
Aeromonas hydrophila SUBSP-HYDROPHILA-ATCC-7966 A0KNU8
Aeromonas salmonicida SUBSP-SALMONICIDA-A449 A4SJB7
Aeromonadales
(Aeromonas)
Vibrio splendidus LGP32 B7VST2
Photobacterium profundum Q6LRZ8
Photobacterium profundum 3TCK
Vibrio parahaemolyticus RIMD-2210633 Q87FT3
Vibrio vulnificus YJ016 Q7MCD1
Vibrio vulnificus CMCP6 Q8D5F3
Vibrionales
(Vibrio, Photobacterium profundum)
Shewanella amazonensis SB2B A1S7J2
Shewanella denitrificans OS217 Q12M19
Shewanella frigidimarina NCIMB-400 Q084Q3
Moritella loihiensis SP-PE36
Shewanella woodyi ATCC-51908 B1KKL1
Shewanella benthica KT99
Shewanella sediminis HAW-EB3 A8FTU9
Shewanella loihica PV-4 A3QF88
Shewanella piezotolerans WP3 B8CPV3
Shewanella halifaxensis HAW-EB4 B0TKY2
Shewanella pealeana ATCC-700345 A8H322
Shewanella oneidensis MR-1 Q8ECS7
Shewanella piezotolerans SP-ANA-3 A0KVA7
Shewanella sediminis SP-MR-4 Q0HKA5
Shewanella sediminis SP-MR-7 Q0HWK7
Shewanella putrefaciens CN-32 A4Y867
Shewanella sediminis SP-W3-18-1 A1RIC1
Shewanella baltica OS155 A3D659
Shewanella baltica OS223 B8EEC4
Shewanella baltica OS185 A6WPZ7
Shewanella baltica OS195 A9KTU0
Alteromonadales
(Shewanella, Moritella loihiensis)
Gammaproteobacteria
Aspergillus fumigatus AF293 Q4WHL2
Aspergillus niger A2QWD1
Yarrowia lipolytica Q6CG13
Candida albicans WO-1 C4YK65
Candida dubliniensis CD36 B9WC51
Debaryomyces hansenii Q6BKS1
Pichia stipitis CBS-6054 A3LYH1
Oryza sativa INDICA-GROUP B8B380
Arabidopsis thaliana Q9M314
Populus trichocarpa
Populus trichocarpa
Chlamydomonas reinhardtii
Micromonas reinhardtii SP-RCC299 C1FGG8
Ostreococcus lucimarinus CCE9901 A4S1Q9
Ostreococcus tauri Q012Y0
Eukaryotes
Deinococcus deserti VCD115
Deinococcus-thermus
Propionibacterium acnes KPA171202 Q6A5T3
Corynebacterium efficiens YS-314 Q8FS10
Corynebacterium glutamicum ATCC-13032 Q8NSS7
Corynebacterium glutamicum R A4QBT3
Saccharopolyspora erythraea NRRL-2338 A4FFA2
Streptomyces griseus SUBSP-GRISEUS-NBRC-13350 B1VQN4
Streptomyces avermitilis MA-4680 Q825N4
Streptomyces coelicolor A3-2- Q9RD72
Kineococcus radiotolerans SRS30216 A6WEX7
Corynebacterium diphtheriae NCTC-13129 Q6NFZ1
Corynebacterium aurimucosum ATCC-700975
Renibacterium salmoninarum ATCC-33209 A9WT82
Arthrobacter aurescens TC1 A1R6F6
Arthrobacter chlorophenolicus A6 B8H7K9
Arthrobacter chlorophenolicus SP-FB24 A0JWN2
Nocardioides radiotolerans SP-JS614 A1SQH1
Clavibacter michiganensis SUBSP-MICHIGANENSIS-NCPPB-382 A5CP43
Clavibacter michiganensis SUBSP-SEPEDONICUS B0RHT0
Marinomonas chejuensis SP-MWYL1 A6VY53
Rhizobium leguminosarum BV-VICIAE-3841 Q1M7N3
Renibacterium salmoninarum ATCC-33209 A9WN22
Bifidobacterium animalis SUBSP-LACTIS-AD011 B8DTF3
Bifidobacterium longum SUBSP-INFANTIS-ATCC-15697 B7GST7
Bifidobacterium adolescentis ATCC-15703 A1A105
Bifidobacterium longum DJO10A B3DSJ6
Bifidobacterium longum NCC2705 Q8G5S5
Gammaproteobacteria
Alphaproteobacteria
Actinobacteria
0.2

Figure S3



Chloroherpeton thalassium B3QUN8
Prosthecochloris aestuarii B4S400
Chlorobaculum parvum B3QQC3
Pelodictyon luteolum Q3B259
Chlorobium phaeobacteroides A1BDL1
Chlorobi
100
63
63
71
Aquifex aeolicus O66459
Hydrogenobaculum aeolicus B4U681
Aquificales
100
Persephonella marina
Sulfurihydrogenibium azorense
100
100
Campylobacter lari
Wolinella succinogenes Q7M7Q1
65
Nitratiruptor profundicola A6Q5D2
100
Sulfurimonas denitrificans Q30P87
Sulfurovum profundicola A6Q6R0
91
100
Thiomicrospira crunogena Q31GY6
Pseudoalteromonas haloplanktis Q3IG05
Psychrobacter arcticus Q4FRT9
Marinobacter aquaeolei A1U3H5
91
Other Proteobacteria
100
Xanthomonas campestris B0RRW7
Methylococcus capsulatus Q606Q4
Nitrosospira multiformis Q2YAB4
Laribacter hongkongensis
Aromatoleum aromaticum Q5P7F3
71
83
100
76
Neorickettsia sennetsu Q2GDE2
Ehrlichia ruminantium Q5FEN4
90
91
Orientia tsutsugamushi B3CRN9
Rickettsia bellii A8GWY3
99
Alphaproteobacteria
100
Magnetospirillum magneticum Q2VZT2
Pelagibacter ubique Q4FPG6
Phenylobacterium zucineum B4R8R5
63
21
Sinorhizobium medicae A6U9Q2
Rhodobacter sphaeroides A4WWC5
Sphingomonas wittichii A5V643
57
100
7
12
Arabidopsis thaliana A7KNF9
Populus trichocarpa
100
Eukaryotes
Mus musculus P00158
Danio rerio Q9MIX8
100
98
Schizosaccharomyces pombe P05501
73
Candida glabrata Q85QA4
97
96
Aspergillus niger Q33798
Ustilago maydis Q0H8X3
100
0.1

Figure S4. KEGG categories

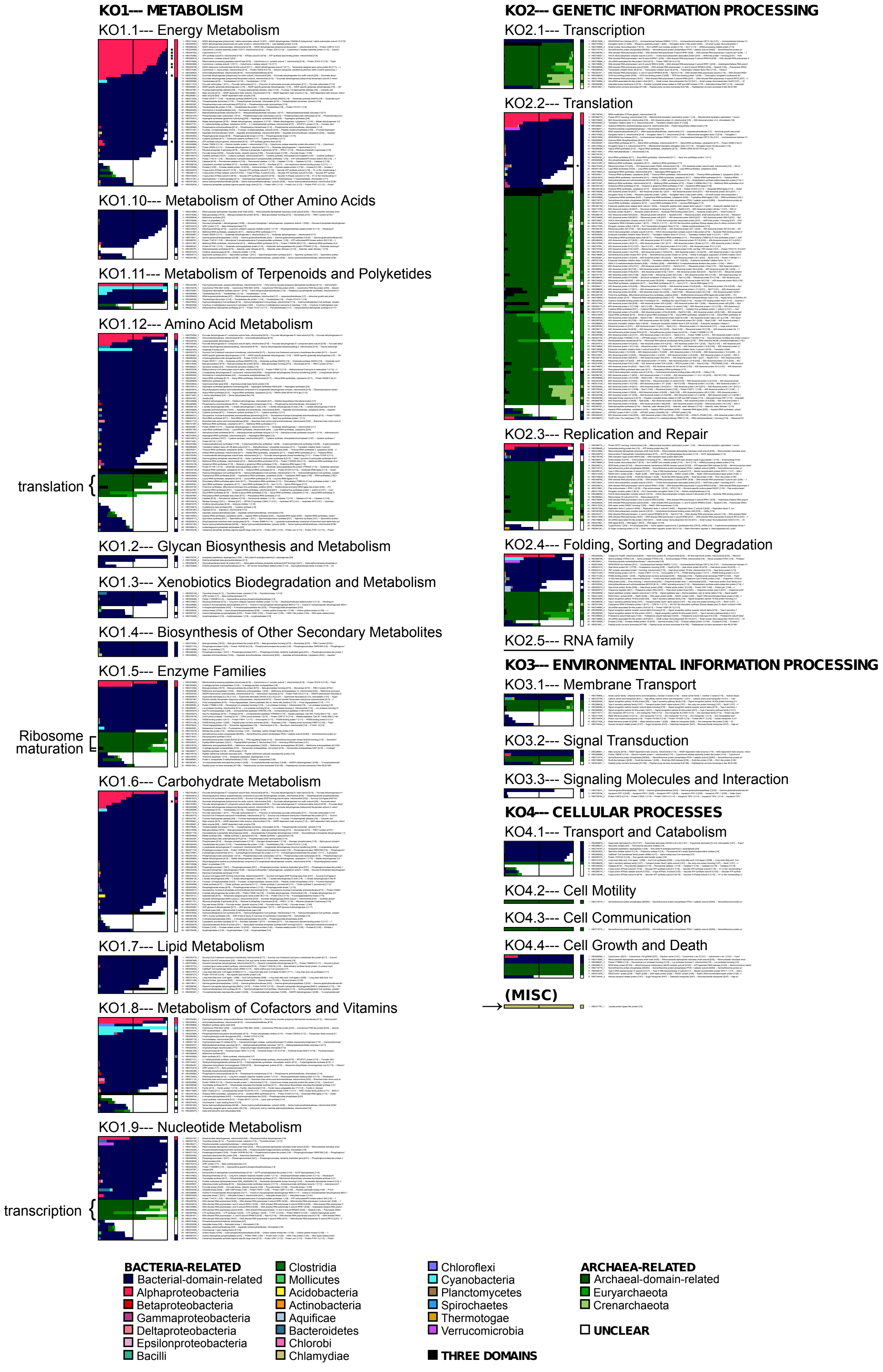


Figure S5

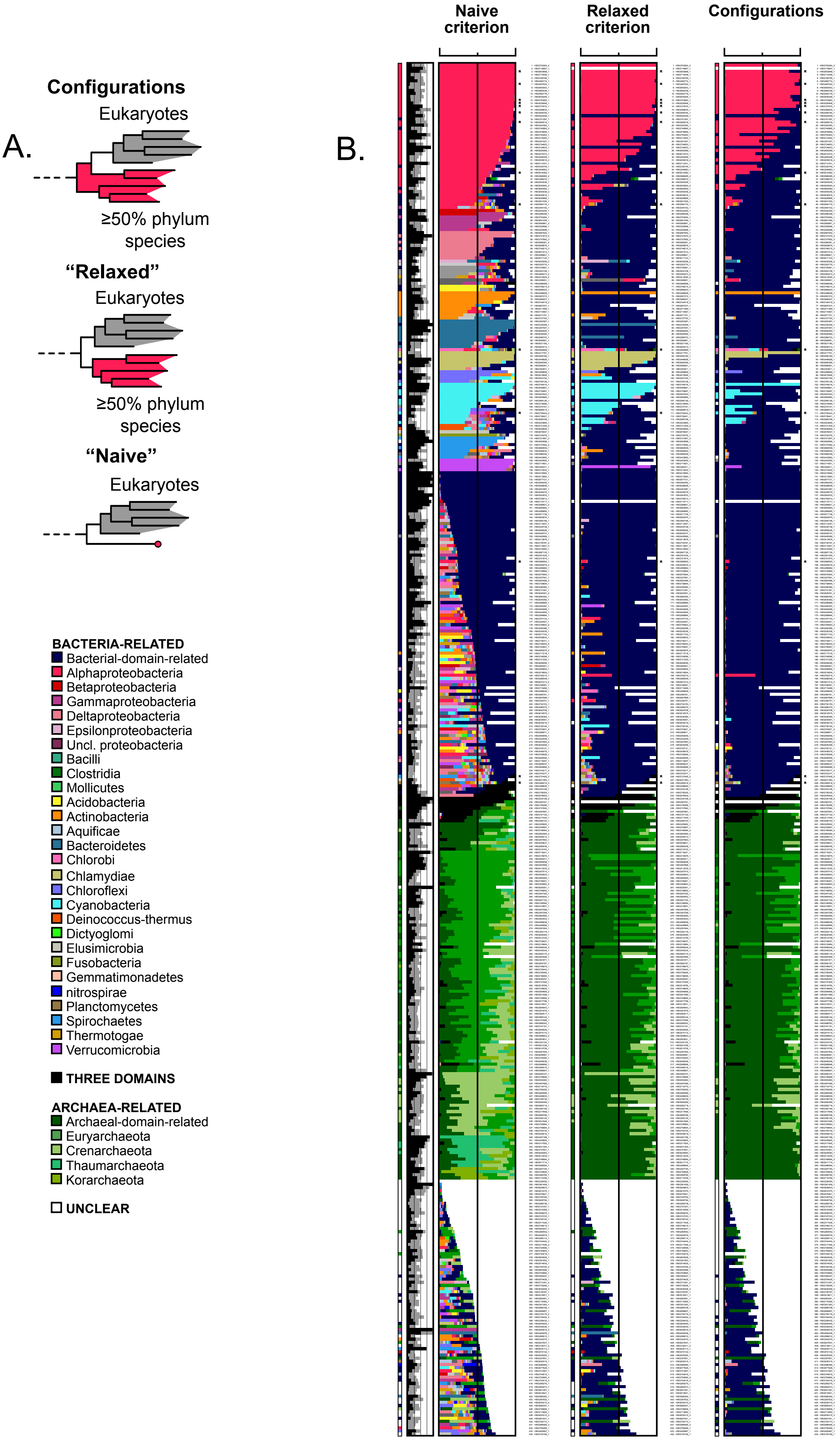

A.
Configurations
Eukaryotes
≥50% phylum species
"Relaxed"
Eukaryotes
≥50% phylum species
"Naive"
Eukaryotes
BACTERIA-RELATED
Bacterial-domain-related
Alphaproteobacteria
Betaproteobacteria
Gammaproteobacteria
Deltaproteobacteria
Epsilonproteobacteria
Uncl. proteobacteria
Bacilli
Clostridia
Mollicutes
Acidobacteria
Actinobacteria
Aquificae
Bacteroidetes
Chlorobi
Chlamydiae
Chloroflexi
Cyanobacteria
Deinococcus-thermus
Dictyoglomi
Elusimicrobia
Fusobacteria
Gemmatimonadetes
nitrospirae
Planctomycetes
Spirochaetes
Thermotogae
Verrucomicrobia
THREE DOMAINS
ARCHAEA-RELATED
Archaeal-domain-related
Euryarchaeota
Crenarchaeota
Thaumarchaeota
Korarchaeota
UNCLEAR
B.
Naive criterion
Relaxed criterion
Configurations

Table S1: List of our 28 near-universal genes and overlap with other studies.

	Yeast sequence	Also used by :		
LECA clade	(Uniprot)	Guy2011	Williams2012 (GI)	Description
HBG299153_1	P25443	COG0098	6321315	40S ribosomal protein S2
HBG304663_2	P36132	COG0533	37362674	Probable tRNA threonylcarbamoyladenosine biosynthesis protein KAE1
HBG307515_1	Q01855			40S ribosomal protein S15
HBG317444_1	P38701			40S ribosomal protein S20
HBG385416_1	P06367	COG0100	10383793	40S ribosomal protein S14-A
HBG397258_1	P33322			H/ACA ribonucleoprotein complex subunit 4
HBG439330_1	P04456			60S ribosomal protein L25
HBG457663_1				Lost in yeast ; Q5VV42 (human) : Threonylcarbamoyladenosine tRNA methylthiotransferase
HBG518785_1	P0CX51	COG0103	9755341	40S ribosomal protein S16-A
HBG562580_1	P20424		6325345	Signal recognition particle subunit SRP54
HBG562580_2	P32916		398366427	Signal recognition particle receptor subunit alpha homolog
HBG568696_1	P05740	COG0091		60S ribosomal protein L17-A
HBG584843_1	P53942			Ribonuclease H2 subunit A (RNase H2 subunit A) (EC 3.1.26.4)
HBG586249_2	P42942			Uncharacterized GTP-binding protein YGR210C
HBG592167_1	P0C0W9	COG0094	6325359	60S ribosomal protein L11-A
HBG594170_2	P0CX53	COG0080	6320781	60S ribosomal protein L12-A
HBG594899_2	P0CX41	COG0093	398364725	60S ribosomal protein L23-A
HBG606796_1	P05738	COG0097		60S ribosomal protein L9-A
HBG635083_1	P39730		6319282	Eukaryotic translation initiation factor 5B
HBG727142_1	P02406			60S ribosomal protein L28
HBG735443_1	P0C0W1	COG0096		40S ribosomal protein S22-A
HBG736296_1	P32905	COG0052	6323077	40S ribosomal protein S0-A
HBG737405_1	P0CX55	COG0099		40S ribosomal protein S18-A
HBG737692_1	P32324		6324707	Elongation factor 2
HBG747181_1	P0CX29	COG0048	6325389	40S ribosomal protein S23-A
HBG748739_1	P0CX45		6322171	60S ribosomal protein L2-A
HBG748975_1	P05750	COG0092	398364505	40S ribosomal protein S3
HBG750455_1	P0CX47	COG0186		40S ribosomal protein S11-A